Machine Learning: Topic Chart

- **Core problems of pattern recognition**
- ***Bayesian decision theory***
- **Support vector machines**
- **Data clustering**
- **Dimension reduction**

Bayesian Decision Theory

The Problem of Statistical Decisions

Task: n **objects have to be partitioned in** $1, \ldots, k$ **classes,** the doubt class $\mathcal{D}$ and the outlier class $\mathcal{O}$.

$\mathcal{D}$: **doubt class** ($\rightarrow$ new measurements required)
$\mathcal{O}$: **outlier class**, definitively none of the classes $1, 2, \ldots, k$

Objects are characterized by feature vectors $X \in \mathcal{X}$, $X \sim \mathbf{P}(X)$ with the probability $\mathbf{P}(X = x)$ of feature values x.

Statistical modeling: Objects represented by data X and classes Y are considered to be random variables, i.e., $(X, Y) \sim \mathbf{P}(X, Y)$.

Conceptually, it is not mandatory to consider class labels as random since they might be induced by legal considerations or conventions.

Structure of the feature space $\mathcal{X}$

- $\mathcal{X} \subset \mathbb{R}^d$
- $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$ with $\mathcal{X}_i \subseteq \mathbb{R}$ or $\mathcal{X}_i$ finite.

Remark: in most situations we can define the feature space as subsets of $\mathbb{R}^d$ or as tuples of real, categorial or ordinal numbers; sometimes we have more complicated data spaces composed of lists, trees or graphs.

Class density / likelihood: $p_y(x) := \mathbf{P}(X = x | Y = y)$ is equal to the probability of a feature value x given a class y.

Parametric Statistics: estimate the parameters of the class densities $p_y(x)$

Non-Parametric Statistics: minimize the empirical risk

Motivation of CLassification

Given are labeled data
$\mathcal{Z} = \{(x_i, y_i) : 1 \leq i \leq n\}$

Questions:

1. What are the class boundaries?
2. What are the class specific densities $p_y(x)$?
3. How many modes do we need to model $p_y(x)$?
4. ...

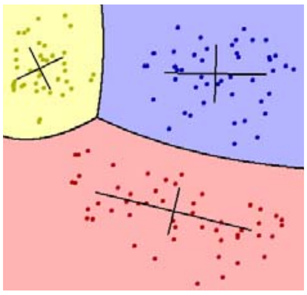


Figure: three Gaussian densities are fitted to the given data samples. The crosses denote the principal axes of the Gaussians.

Thomas Bayes and his Terminology

The **State of Nature** is modelled as a random variable!

prior:	$\mathbf{P}\{\text{model}\}$
likelihood:	$\mathbf{P}\{\text{data}\|\text{model}\}$
posterior:	$\mathbf{P}\{\text{model}\|\text{data}\}$
evidence:	$\mathbf{P}\{\text{data}\}$

Bayes Rule: $\mathbf{P}\{\text{model}|\text{data}\} = \dfrac{\mathbf{P}\{\text{data}|\text{model}\}\mathbf{P}\{\text{model}\}}{\mathbf{P}\{\text{data}\}}$

Ronald A. Fisher and Frequentism

Fisher, Ronald Aylmer (1890-1962): founder of frequentist statistics together with Jerzey Neyman & Karl Pearson.

photo A Barrington-Brown (c) R A Fisher Memorial Trust.

British mathematician and biologist who invented revolutionary techniques for applying statistics to natural sciences.

Maximum likelihood method

Fisher information: a measure for the information content of densities.

Sampling theory

Hypothesis testing

Bayesianism vs. Frequentist Inference[1]

Bayesianism is the philosophical tenet that the mathematical theory of probability applies to the degree of plausibility of statements, or to the degree of belief of rational agents in the truth of statements; together with Bayes theorem, it becomes Bayesian inference. The Bayesian interpretation of probability allows probabilities assigned to random events, but also allows the assignment of probabilities to any other kind of statement.

Bayesians assign probabilities to any statement, even when no random process is involved, as a way to represent its plausibility. As such, the scope of Bayesian inquiries include the scope of frequentist inquiries.

The limiting relative frequency of an event over a long series of trials is the conceptual foundation of the frequency interpretation of probability.

Frequentism rejects degree-of-belief interpretations of mathematical probability as in Bayesianism, and assigns probabilities only to random events according to their Relative frequencies of occurrence.

[1] see `http://encyclopedia.thefreedictionary.com/`

Bayes Rule for Known Densities and Parameters

Classifier:

$$\hat{c} : \mathcal{X} \to \{1, \dots, k, \mathcal{D}\}$$

The assignment function $\hat{c}$ maps the feature space $\mathcal{X}$ to the set of classes $\{1, \dots, k, \mathcal{D}\}$. (Outliers are neglected)

Quality of the classifier: **expected risk**

$$\mathcal{R}(\hat{c}) = \sum_{y \le k} \mathbf{P}(y) \mathsf{E}_x \left[\mathbb{I}_{\{\hat{c}(x) \ne y\}} | Y = y \right] + \text{terms from } \mathcal{D}$$

Remark: The rational behind this choice comes from gambling. If we bet on a particular outcome of our experiment and our gain is measured by how often we assign the measurements to the correct class then classifier with minimal expected risk will win *on average* against any other classification rule ("Dutch books")!

Loss Function: $L(y, z)$ denotes the loss for the decision z if class y is correct.

0-1 loss: all classes are treated the same!

$$L^{0-1}(y, z) = \begin{cases} 0 & \text{if } z = y \text{ (correct decision)} \\ 1 & \text{if } z \ne y \text{ and } z \ne \mathcal{D} \text{ (wrong decision)} \\ d & \text{if } z = \mathcal{D} \text{ (no decision)} \end{cases}$$

- weighted classification costs $L(y, z) \in \mathbb{R}^+$ are frequently used, e.g. in medicine
  classification costs can also be asymmetric, that means $L(y, z) \ne L(z, y)$
  ($(z, y) \sim$ (stomach cancer, stress related stomach problem).

Risk function of the classifier is the expected loss/costs:

$$\begin{aligned} \mathcal{R}(\hat{c}, y) &= \mathsf{E}_x \left[L(y, \hat{c}(x)) | Y = y \right] \\ &= \sum_{z \le k} \Big(L(y, z) \mathbf{P}\{\hat{c}(x) = z | Y = y\} \\ &\quad + L(y, \mathcal{D}) \mathbf{P}\{\hat{c}(x) = \mathcal{D} | Y = y\} \Big) \\ &= \underbrace{\mathbf{P}\{\hat{c}(x) \ne y \wedge \hat{c}(x) \ne \mathcal{D} | Y = y\}}_{\text{pmc}(y) \text{ misclassification probability}} + \underbrace{d \cdot \mathbf{P}\{\hat{c}(x) = \mathcal{D} | Y = y\}}_{\text{pd}(y) \text{ doubt probability}} \end{aligned}$$

Total risk of the classifier: ($\pi_y := \mathbf{P}(Y = y)$)

$$\mathcal{R}(\hat{c}) = \sum_{z \le k} \pi_z \, \text{pmc}(z) + d \sum_{z \le k} \pi_z \, \text{pd}(z) = \mathsf{E}_C \left[\mathcal{R}(\hat{c}, C) \right]$$

Asymptotic average loss

$$\lim_{n \to \infty} \frac{1}{n} \sum_{j \le n} L(c_j, \hat{c}(x_j)) = \lim_{n \to \infty} \hat{\mathcal{R}}(\hat{c}) = \mathcal{R}(\hat{c}),$$

where $\{(x_j, c_j) | 1 \le j \le n\}$ is a random sample set of size n.

This formula can be interpreted as the expected loss with empirical distribution as probability model.

Posterior class probability : Let

$$p(y|x) \equiv \mathbf{P}\{Y = y | X = x\} = \frac{\pi_y p_y(x)}{\sum_z \pi_z p_z(x)}$$

be the posterior of the class y given $X = x$.

(The 'Partition of One" $\pi_y p_y(x) / \sum_z \pi_z p_z(x)$ results from the normalization $\sum_z p(z|x) = 1$.)

Bayes Optimal Classifier

Theorem 1 *The classification rule which minimizes the total risk for* $0 - 1$ *loss is*

$$c(x) = \begin{cases} y & \textit{if} \quad p(y|x) = \max_{z \le k} p(z|x) > 1 - d, \\ \mathcal{D} & \textit{if} \quad p(y|x) \le 1 - d \quad \forall y. \end{cases}$$

Generalization to arbitrary loss functions

$$c(x) = \begin{cases} y & \text{if} \quad \sum_z L(z, y) p(z|x) = \min_{\rho \le k} \sum_z L(z, \rho) p(z|x) \le d, \\ \mathcal{D} & \text{else}\,. \end{cases}$$

Bayes classifier: Select the class with highest $\pi_y p_y(x)$ value if it exceeds the costs for not making a decision, i.e., $\pi_y p_y(x) > (1 - d) p(x)$.

Proof: Calculate the total expected loss $\mathcal{R}(\hat{c})$

$$\begin{aligned}\mathcal{R}(\hat{c}) &= \mathsf{E}_X\left[\mathsf{E}_Y\left[L^{0-1}(Y,\hat{c}(x))|X=x\right]\right]\\ &= \int_{\mathcal{X}} \mathsf{E}_Y\left[L^{0-1}(Y,\hat{c}(x))|X=x\right]p(x)dx \text{ with } p(x)=\sum_{z\le k}\pi_z p_z(x)\end{aligned}$$

Minimize the conditional expectation value since it depends only on $\hat{c}$.

$$\begin{aligned}\hat{c}(x) &= \text{argmin}_{\tilde{c}\in\{1,\dots,k,\mathcal{D}\}}\mathsf{E}\left[L^{0-1}(Y,\tilde{c})|X=x\right]\\ &= \text{argmin}_{\tilde{c}\in\{1,\dots,k,\mathcal{D}\}}\sum_{z\le k}L^{0-1}(z,\tilde{c})p(z|x)\\ &= \begin{cases}\text{argmin}_{\tilde{c}\in\{1,\dots,k\}}(1-p(\tilde{c}|x)) & \text{if } d>\min_c(1-p(c|x))\\ \mathcal{D} & \text{else}\end{cases}\\ &= \begin{cases}\text{argmax}_{\tilde{c}\in\{1,\dots,k\}}p(\tilde{c}|x) & \text{if } 1-d<\max_c p(c|x)\\ \mathcal{D} & \text{else}\end{cases} \quad \square\end{aligned}$$

Outliers

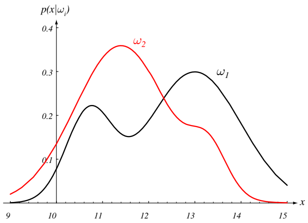
- Modeling by an outlier class $\pi_{\mathcal{O}}$ with $p_{\mathcal{O}}(x)$
- **"Novelty Detection"**: Classify a measurement as an outlier if
$$\pi_{\mathcal{O}}p_{\mathcal{O}}(x) \ge \max\left\{(1-d)p(x), \max_z \pi_z p_z(x)\right\}$$
- The outlier concept causes conceptual problems and it does not fit to the statistical decision theory since outliers indicate an erroneous or incomplete specification of the statistical model!
- The outlier class is often modeled by a uniform distribution.
Attention: Normalization of uniform distribution does not exist in many feature spaces!

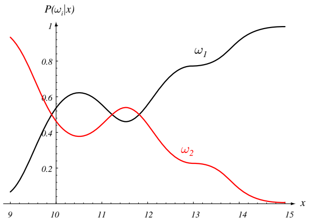
  $\Longrightarrow$ Limit the support of the measurement space or put a (Gaussian) measure on it!

Class Conditional Densities and Posteriors for 2 Classes

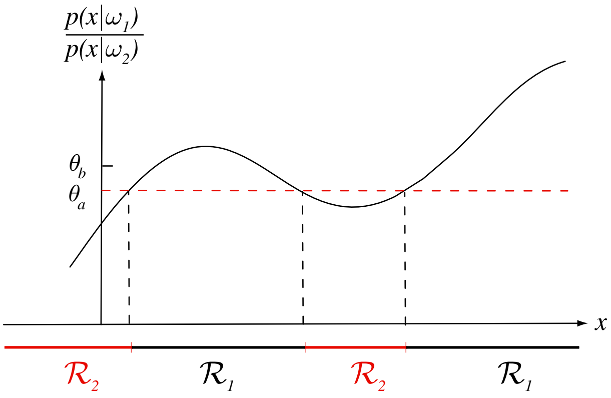
Class-conditional probability density function

Posterior probabilities for priors $\mathbf{P}(y_1)=\frac{2}{3}, \mathbf{P}(y_2)=\frac{1}{3}$.

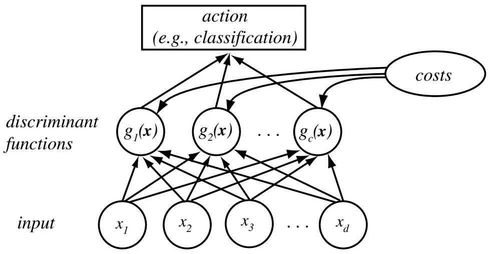




Likelihood Ratio for 2 Class Example

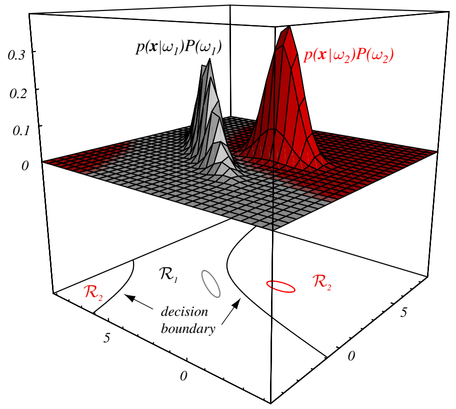


Discriminant Functions g_l

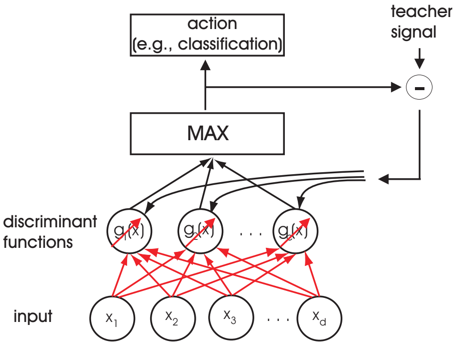


- Discriminant function: $g_z(x) = \mathbf{P}\{Y=y|X=x\}$
- Class decision: $g_y(x) > g_z(x)\ \forall z \ne y \ \Rightarrow$ class y.
- Different discriminant functions can yield the same decision: $\tilde{g}_y(x) = \log \mathbf{P}\{x|y\} + \log \pi_y$; minimize implementation problems!

Example for Discriminant Functions



Adaptation of Discriminant Functions g_l



action (e.g., classification)

teacher signal

MAX

discriminant functions

input

The red connections (weights) are adapted in such a way that the teacher signal is imitated by the discriminant function.

Example Discriminant Functions: Normal Distributions

The *Likelihood* of class y is Gaussian distributed.

$$p_y(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_y|}} \exp\left(-\frac{1}{2}(x-\mu_y)^T\Sigma_y^{-1}(x-\mu_y)\right)$$

$\boxed{\textbf{Special case: } \Sigma_y = \sigma^2\mathbb{I}}$

$$\begin{aligned} g_y(x) &= \log p_y(x) + \log \pi_y \\ &= -\frac{1}{2\sigma^2}\|x-\mu_y\|^2 + \log\pi_y + const. \end{aligned}$$

⇒ Decision surface between class z and y:

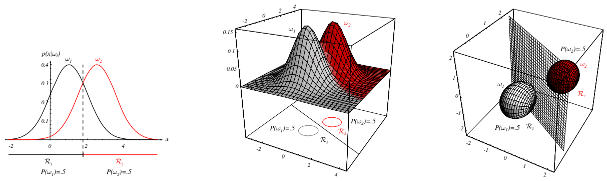
$$\begin{aligned} -\frac{1}{2\sigma^2}\|x-\mu_z\|^2 + \log\pi_z &= -\frac{1}{2\sigma^2}\|x-\mu_y\|^2 + \log\pi_y \\ -\|x\|^2 + 2x\cdot\mu_z - \|\mu_z\|^2 + 2\sigma^2\log\pi_z &= -\|x\|^2 + 2x\cdot\mu_y - \|\mu_y\|^2 + 2\sigma^2\log\pi_y \end{aligned}$$

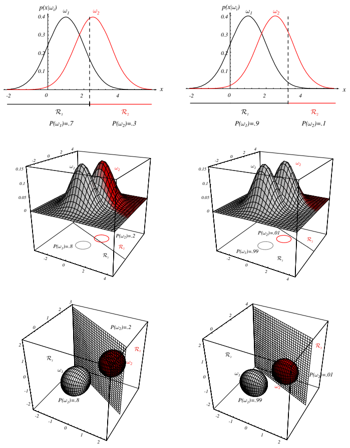
$$\Rightarrow \qquad 2x\cdot(\mu_z-\mu_y) - \|\mu_z\|^2 + \|\mu_y\|^2 + 2\sigma^2\log\frac{\pi_z}{\pi_y} = 0$$

Linear decision rule: $\boxed{w^T(x-x_0) = 0}$

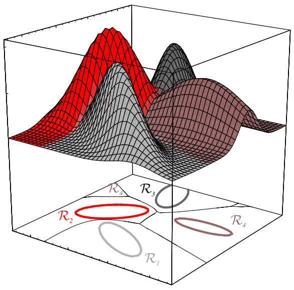
with $\quad w = \mu_z - \mu_y \qquad x_0 = \frac{1}{2}(\mu_z+\mu_y) - \frac{\sigma^2(\mu_z-\mu_y)}{\|\mu_z-\mu_y\|^2}\log\frac{\pi_z}{\pi_y}$

Decision Surface for Gaussians in 1,2,3 Dimensions





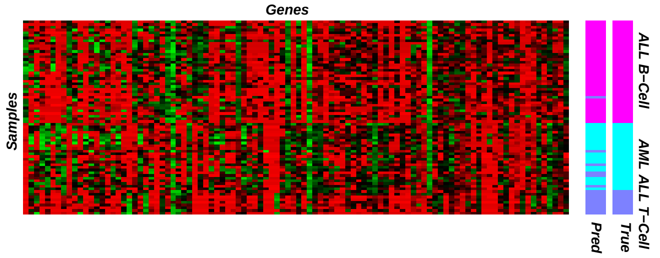
Multi Class Case



Decision regions for four Gaussian distributions. Even for such a small number of classes the discriminant functions show a complex form.

Example: Gene Expression Data

The expression of genes is measured for various patients. The expression profiles provide information of the metabolic state of the cells, meaning that they could be used as indicators for disease classes. Each patient is represented as a vector in a high dimensional (≈ 10000) space with Gaussian class distribution.



Parametric Models for Class Densities

If we would know the prior probabilities and the class conditional probabilities then we could calculate the optimal classifier. But we don't!

Task: Estimate $p(c|x;\theta)$ from samples $\mathcal{Z} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ for classification.

Data are sorted according to their classes: $\mathcal{X}_y = \{X_{1y}, \dots, X_{n_y,y}\}$ where $X_{iy} \sim \mathbf{P}\{X|Y = y; \theta_y\}$

Question: How can we use the information in samples to estimate θ_y?

Assumption: classes can be separated and treated independently! $\mathcal{X}_y$ **is not informative w.r.t.** $\theta_\alpha,\ \alpha \neq y$

Maximum Likelihood Estimation Theory

Likelihood of the data set: $\mathbf{P}\{\mathcal{X}_y|\theta_y\} = \prod_{i \le n_y} p(x_{iy}|\theta_y)$

Estimation principle: Select the parameter $\hat{\theta}_y$ which maximizes the likelihood, that means

$$\hat{\theta}_y = \arg\max_{\theta_y} \mathbf{P}\{\mathcal{X}_y|\theta_y\}$$

Procedure: Find the extreme value of the log-likelihood function

$$\begin{aligned} \nabla_{\theta_y} \log \mathbf{P}\{\mathcal{X}|\theta_y\} &= 0 \\ \frac{\partial}{\partial \theta_y} \sum_{i \le n} \log p(x_i|\theta_y) &= 0 \end{aligned}$$

Remark

Bias of an estimator: bias$(\hat{\theta}_n) = \mathsf{E}\{\hat{\theta}_n\} - \theta$.

Consistent estimator: A point estimator $\hat{\theta}_n$ of a parameter θ is consistent if $\hat{\theta}_n \xrightarrow{\mathbf{P}} \theta$.

Asymptotic Normality of Maximum Likelihood estimates: $(\hat{\theta}_n - \theta)/\sqrt{\mathsf{V}\{\hat{\theta}_n\}} \rightsquigarrow \mathcal{N}(0, 1)$.

Alternative to ML class density estimation: **discriminative learning** by maximizing the a posteriori distribution $\mathbf{P}\{\theta_y|\mathcal{X}_y\}$ (details of the density do not have to be modelled since they might not influence the posterior)

Example: Multivariate Normal Distribution

Expectation values of a normal distribution and its estimation:
Class index has been omitted for legibility reasons ($\theta_y \to \theta$).

$$\begin{aligned} \log p(x_i|\theta) &= -\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu) - \frac{d}{2} \log 2\pi - \frac{1}{2} \log|\Sigma| \\ \frac{\partial}{\partial \mu} \sum_{i \le n} \log p(x_i|\theta) &= \frac{1}{2} \sum_{i \le n} \Sigma^{-1}(x_i - \mu) + \frac{1}{2} \sum_{i \le n} \left((x_i - \mu)\Sigma^{-1}\right)^T = 0 \\ \Sigma^{-1} \sum_{i \le n} (x_i - \mu) &= 0 \quad \Rightarrow \quad \hat{\mu}_n = \frac{1}{n} \sum_i x_i \qquad \text{estimator for } \mu \end{aligned}$$

Average value formula results from the quadratic form.

Unbiasedness: $\mathsf{E}[\hat{\mu}_n] = \frac{1}{n} \sum_{i \le n} \mathsf{E} x_i = \mathsf{E}[x] = \mu$

ML estimation of the variance (1d case)

$$\begin{aligned} \frac{\partial}{\partial \sigma^2} \sum_{i \le n} \log p(x_i|\theta) &= -\frac{\partial}{\partial \sigma^2} \sum_{i \le n} \frac{1}{\sigma^2} \|x_i - \mu\|^2 - \frac{n}{2} \log(2\pi\sigma^2) \\ &= \frac{1}{2} \sum_{i \le n} \sigma^{-4} \|x_i - \mu\|^2 - \frac{n}{2} \sigma^{-2} \ = \ 0 \\ \Rightarrow \quad \hat{\sigma}_n^2 &= \frac{1}{n} \sum_{i \le n} \|x_i - \mu\|^2 \end{aligned}$$

Multivariate case $\hat{\Sigma}_n = \frac{1}{n} \sum_{i \le n} (x_i - \mu)(x_i - \mu)^T$

$\hat{\Sigma}_n$ is biased, e.g., $\mathsf{E}\hat{\Sigma}_n \neq \Sigma$, if μ is unknown.